Section II. Technical Notes

The data on doctoral sciences and engineers contained in this report come from the 1999 Survey of Doctorate Recipients (SDR),[1] which is a longitudinal panel survey of individuals who have received their doctorates in the sciences or engineering (S&E). Since the 1970s, this study has been conducted every two years for the National Science Foundation (NSF) and other federal sponsors.[2]

The U.S. Census Bureau conducted the survey for the NSF in 1999. Data collected in the SDR are part of the Sciences and Engineers Statistical Data System (SESTAT), surveys that are sponsored and maintained by NSF. Additional data on education and demographic information in the SDR come from the Survey of Earned Doctorates (SED), an ongoing annual census of research doctorates earned in the United States since 1920, which forms the Doctorate Records File (DRF).

The Sampling Frame and Target Population

The sampling frame for the 1999 SDR was compiled from the DRF to include individuals who:

1. Had earned a doctoral degree from a U.S. college or university in a S&E field[3]

2. Were U.S. citizens or, if non-U.S. citizens, indicated they had plans to remain in the United States after degree award

3. Were under 76 years of age

The 1999 frame consisted of the 1997 SDR sample supplemented with new S&E doctorate graduates who had earned their doctoral degrees since the 1997 survey and who met the conditions listed above. Those who were carried over from 1997 but had attained the age of 76 (or were deceased) were deleted from the frame.

The survey had two additional eligibility criteria for the survey target population. The sampled member must be a resident of the United States and not institutionalized as of the survey reference date.

Sample Design

In 1999, the SDR sample size was 40,000. The total sample was selected from three groups:

- Old cohort cases with doctoral degrees earned prior to July 1, 1992

- Nearly new cohort cases with doctoral degrees earned between July 1, 1992 and June 30, 1996

- New cohort cases with doctoral degrees earned between July 1, 1996 and June 20, 1998

The goals of the 1999 SDR sample design included the following:

- Reduce the variation in the sampling weights of the old and nearly new cohorts

- Allocate the sample so the variance of overall population estimates are minimized

- Allocate the sample so the sampling rate of the new cohort is at least 15 percent higher than that of the old cohort

- Allocate the sample so the sampling rate of the nearly new cohort is at least 10 percent higher than that of the old cohort

- Adjust the sample allocation if any large stratum receives a disproportionate amount of sample

To ensure that the sampling rate of the new cohort was at least 15 percent higher than that of the old cohort, 4,000 of the total sample was from the new cohort group. The remaining 36,000 sample cases were then divided so that the nearly new cohort would have a 10 percent higher sample allocation than the old cohort.

The basic sampling design was a stratified design where strata were defined by 15 broad fields of study, 2 genders, and an 8-category "demographic group" variable combining race/ethnicity, disability status, and citizenship

[1] The discussions presented here are partly from the 1999 Survey of Doctorate Recipients Methodology Report (Census Bureau, 2002).

[2] In 1999, the National Institutes of Health co-sponsored the SDR with NSF. In previous rounds, the Department of Energy and the National Endowment for the Humanities co-sponsored the survey.

[3] See Appendix A for a list of the science and engineering fields included in the 1999 SDR sampling frame.

status. The sample cases were combined in the multi-way cross of the stratification variables. The sample allocation to select the cases from each stratum followed a seven-step process. For strata where the allocated sample size was equal to the frame size, all cases were selected for the sample. For all other strata, sample cases were selected using the probability-proportional-to-size (PPS) selection method separately for each cohort group (with the sampling weights as the size measure).

The overall sampling rate was about 1 in 16 (6.2 percent) in the 1999 SDR, applied to an estimated science and engineering doctoral population of 650,300. However, sampling rates varied considerably within and among the strata.

Survey Content

The 1999 SDR still retained the questionnaire design changes that were implemented in 1993. A large set of core data items is conveyed from year to year to enable trend comparisons. Each survey year, a different set of module questions on special topics of interest are included. For example, the 1995 SDR questionnaire had a postdoc module and the 1997 had special modules on alternative work arrangements, job security concerns, and recent doctorates' initial career experiences. No special module was introduced in the 1999 questionnaire except for the retention of a few recent doctorate module questions from 1997, such as first career path job and doctoral training experiences.

Data Collection

The 1999 SDR data collection consisted of two phases: a self-administered mail survey, followed by computer-assisted telephone interviewing (CATI) of a sample of the nonrespondents to the mail survey. The mail survey consisted of an advance letter and the several waves of a personalized mailing package, with a reminder postcard between the first and second questionnaire mailing. The advance letter was sent in May 1999, followed by the first mailing a week later. The second mailing via USPS priority mail was sent in July 1999. The CATI follow-up of mail nonrespondents ended in February 2000.

Response Rates

The overall unweighted response rate for the 1999 SDR was 81.5 percent. The response to the mail phase of the survey was about 70 percent. The response rate to the CATI phase was about 43 percent. The overall weighted response rate was about 82 percent (weighted response divided by the weighted sample cases).

Data Preparation

Data preparation for the 1999 SDR consisted of clerical, keying, and coding operations performed manually by the Census National Processing Center (NPC) and the computer operations performed by the Census Demographic Surveys Division (DSD). Data preparation began in May 1999 when the first mail questionnaires were returned to the NPC and continued through October 2000 when the DSD delivered the SESTAT formatted, edited, imputed data file to the NSF.

As the mail questionnaires were received, they were checked into the tracking system. The mail-returned questionnaires that had one or more entries were clerically edited for data entry preparation. The clerical edit was limited to simple edits such as correcting illegible entries; rounding fractions to the closest whole number; verifying that city, state, and country entries were in the correct location.

Clerically edited questionnaires were grouped into batches, keyed, and verified using the Key Entry III (KE III) system. The KE III system generated a keying report to track the status of cases through the keying operation. As part of quality control procedures, 5 percent verification was performed of all keyed questionnaires. For some questionnaire items (F9 Birthdate, F13/F14/F16 Contact information), a 100 percent verification of questionnaire items was performed.

NPC transmitted the keyed questionnaire data on a regular basis during the data collection phase to the DSD. DSD performed computer editing to identify cases with missing critical items (A1/A2 labor force status, A6/A21 job codes, F6 resident status in U.S., F9 birthdate) and generated Telephone Follow-up sheets. Telephone call-backs were made to obtain a response to these critical items; otherwise they were considered as incomplete responses. Whenever these callbacks were made, every attempt was made to obtain responses to other missing important data items (A7 FT/PT status, A15/A17 employment sector, A18/A19 type of educational institution, A26 job start date, A30/A31 work activities, and F14 future contact information).

Since the DSD collected data in mail and CATI, the data sets were merged into one data set. The coding operation involved special coding of occupation and education codes, other specify coding, state and country coding and IPEDS coding. For special coding of occupation, the respondent's occupational data were reviewed along with other work-related data from the questionnaire by specially trained coders to "correct" known respondent self-reporting problems to obtain the "best" occupation codes. The education code for a newly earned degree was assigned strictly based on the degree field verbatim.

The "other specify" responses were backcoded to existing response categories using the SESTAT other specify coding guidelines. Employer location (A11), Degreed school location (D6) and Country of citizenship (F8) were assigned the appropriate three-digit FIPS state/country code. The Integrated Postsecondary Education Data System (IPEDS) was used to assign codes for the employers (A11) that are postsecondary institutions and for the newly earned degree school (D6).

A detailed edit specification was developed from the SESTAT edit guidelines to perform further computer editing of multiple values to "Mark One" questions, skip errors, range errors, internal inconsistencies, cross-year inconsistencies. Basic frequency distributions of all survey items showed item nonresponse rates to be generally less than 3 percent. Nonresponse to a few questions deemed somewhat sensitive, such as household income, was around 6.2 percent.

To compensate for item nonresponse, data not reported by the respondents as well as responses of "refused" or "don't know" were imputed. Imputation is a process for treating missing data. Imputation methods are used when answers to questions are blank or not usable. Two imputation methods were used: (1) logical imputation, and (2) hot deck imputation. For logical imputation, either the respondent's answers to related questions determined what the missing value had to be, or the respondent's answer to the same question in the prior survey round was substituted for the missing value. The latter approach of using the historical data is often called "cold deck" imputation. Cold deck imputation is useful for variables that are static, such as place of birth or gender. When logical imputation was used, it was employed before hot deck imputation.

In hot deck imputation, a donor case is selected from the current round of respondents by matching related variables. The donor case's response is used as a proxy for the recipient's missing variable. Hot deck imputation is the method of choice for variables that may change over time, such as employment characteristics. Hot deck is preferable to model-based imputation in this application because it easily preserves correlation among variables and maintains the valid response ranges for categorical variables.

Weighting and Estimation

To enable weighted analyses of the 1999 SDR data, a sample weight was calculated for every person in the sample. The primary purpose of the weights is to create representative estimates by adjusting for unequal probabilities of selection. The second purpose is to adjust for the effects of nonresponse without increasing the variance. Informally, a sampling weight approximates the number of persons in the Ph.D. population that a sampled person represents. A main goal of this weighting plan is to produce final weights that reduce the nonresponse bias in our survey estimates, without increasing the variance.

The weights were calculated in several stages. The first stage was the calculation of base weights that account for the sample design. A base weight is the inverse of the probability of selection in the SDR sample. For cases selected with certainty, the 1999 SDR base weight is equal to the 1999 SDR initial weight. For all other cases, the 1999 SDR base weight is greater than the initial weight. This increase reflects an adjustment for cases not selected for the sample.

From the 1999 SDR base weights, the production of the 1999 SDR final weights involved four main steps:

- Adjustment for duplicate, frame ineligible, and never earned doctorate cases
- Calculation of the 1999 SDR control totals
- Calculation of the 1999 SDR noninterview weights
- Calculation of the 1999 SDR final weights

Raking ratio adjustment was used to control the 1999 SDR sample back to the 1999 SDR population totals. The purpose of this adjustment is twofold:

- To decrease the sampling variability
- To account for changes in the final weights due to changes in the eligible sampling frame

RELIABILITY

Because the estimates produced from this survey are based on a sample, they may vary from those that would have been obtained if all members of the target population had been surveyed (using the same questionnaire and data collection methods). Two types of error are possible when population estimates are derived from any sample survey: sampling error and nonsampling error. By looking at these errors, it is possible to estimate the accuracy and precision of the survey results.

Sampling error is the variation that occurs by chance because a sample, rather than the entire population, is surveyed. The particular sample that was used to estimate the 1999 population of science and engineering doctorates in the United States was one of a large number of samples that could have been selected using the same sample design and size. Estimates based on each of these samples would have differed. Thus, one should be particularly careful when interpreting results based on a relatively small number of cases or on small differences between the estimates.

Due to the large amount of data collected in the SDR, it is not practical to directly calculate variance estimates for every survey estimate. Instead, generalized variance functions were developed to model the variance estimates for certain characteristics. Parameters derived from these generalized variance functions approximate variance estimates for all survey items. As a result, these sampling errors provide an indication of the order of magnitude of a sampling error rather than a precise sampling error for any specific item.

The variances on the survey estimates were calculated by the successive difference replication method. This replication method was used to first calculate a small number of variance estimates, which were then used to estimate the parameters of the generalized variance function. An one-parameter model was used to calculate the generalized variance parameters which were estimated using an iterative weighted least square procedure.

Since many of the SDR estimates of interest consist of small populations such as estimates of Hispanic sciences or black engineers, the finite population correction factor was consistently applied to all the variance estimates.

Different generalized variance functions were used to estimate standard errors associated with a broader range of totals and percentages. The *a* and *b* parameters were calculated for each of the demographic groups and fields of study shown in Appendix C. The *a* and *b* parameters can be used to approximate standard errors for the S&E doctoral population overall, for broad field groupings used by NSF, and for selected subgroups of analytic interest.

STANDARD ERROR OF ESTIMATED NUMBERS

To calculate the desired standard errors on numbers, let *X* denote the estimated number. The standard error can be approximated using the appropriate values of a and b along with the following formula for standard errors of totals:

$$\mathrm{SE}(X) = [aX^2 + bX]^{1/2} \qquad (1)$$

When calculating standard errors for numbers from tabulations involving different characteristics, use the set of parameters for the characteristic which will give the largest standard error.

Illustration

Suppose an estimated 2,770 females with a doctorate in the biological sciences were reported as working in the Federal Government in 1999.

Use the appropriate generalized variance parameters from Appendix C to get:

Survey estimate *X*	=	2,770
a parameter	=	-0.000085
b parameter	=	13.0631

Use formula (1) to approximate the standard error on the estimated number of 2,770 as:

$$\mathrm{SE}(X) = [(-0.000085 \times 2{,}770^2) + (13.0631 \times 2{,}770)]^{1/2}$$
$$= 189$$

The 95% confidence interval is calculated using the following formula:

$$95\%\ \mathrm{CI} = X \pm 1.96 \times \mathrm{SE}(X) \qquad (2)$$

where

X is the survey estimate of interest, and SE(*X*) is the estimated standard error for the survey estimate of interest.

Using formula (2) above, the 95% confidence interval is:

2,770 ± 1.96 × 189 or 2,770 ± 370

Therefore, the 95% confidence interval has the following limits:

Lower limit	=	2,401
Upper limit	=	3,139

So we can say with 95% confidence that the number of females with biological sciences doctorates working in the Federal Government in 1999 is estimated to be between 2,401 and 3,139.

Standard Error of Estimated Percentages

To calculate the standard errors on percentages, let *p* equal the percentage possessing the specific characteristic and *X* and *Y* represent the numerator and denominator, respectively, of the ratio that yields the observed percentage. The standard error of a percentage may be approximated using the formula:

$$\mathrm{SE}(p) = p(\{[\mathrm{SE}(X)^2]/X^2\}) - (\{[\mathrm{SE}(Y)]^2/Y^2\})^{1/2} \qquad (3)$$

where

X and *Y* are survey estimates of interest, SE(*X*) and SE(*Y*) are the corresponding standard error estimates derived using formula (1), and *p* is the estimated percentage ($p = (X/Y) \times 100$).

Illustration

Suppose an estimated 2,770 of the 8,870 biological sciences doctorates working in the Federal Government are women. Therefore, the estimated percentage of biological sciences doctorates working in the Federal Government who are women is 31.2%.

Use formula (1) and the appropriate parameters from Appendix C, to get:

	X	***Y***	***p***
Survey estimate	2,770	8,870	31.2%
a parameter	-0.000085	-0.000092	——
b parameter	13.0631	16.8031	——
Standard error	189	377	

Insert the above numbers into formula (3) to approximate the standard error on the estimate of 31.2% as:

$$\mathrm{SE}(p) = 31.2\ [(189^2/2{,}770^2) - (377^2/8{,}879^2)]^{1/2} = 1.7\%$$

Using formula (2), the 95% confidence interval is:

31.2% ± 1.96 × 1.7% or 31.2% ± 3.3%

Therefore, the 95% confidence interval has the following limits:

Lower limit	=	27.9%
Upper limit	=	34.5%

Standard Error of a Difference

To calculate the standard errors of the difference between two sample estimates, let *X* and *Y* represent two estimates of interest and SE(*X*) and SE(*Y*) the corresponding standard error estimates derived using formula (1).

$$\mathrm{SE}(X - Y) = \{[(\mathrm{SE}(X)]^2 + [\mathrm{SE}(Y)]^2\}^{½} \qquad (4)$$

The estimates can be numbers, percentages, ratios, etc. This will represent the actual standard error quite accurately for the difference between estimates of the same characteristic in two different areas or for the difference between separate and uncorrelated characteristics in the same area.

Illustration

In 1999, suppose there were an estimated 6,100 male and 2,770 female biological sciences doctorates. The apparent difference between the estimated number of male and female biological sciences doctorates is 3,330.

Use the appropriate parameters from Appendix C and formula (1) to get:

	X	***Y***	**Difference**
Survey estimate	6,100	2,770	3,330
a parameter	-0.000092	-0.000085	——
b parameter	16.8031	13.0631	——
Standard error	315	189	

The standard error of the difference is calculated using formula (4):

$$SE(X - Y) = ([315^2 + 189^2])^{1/2} = 367$$

The 95% confidence interval is calculated as 3,330 ± 1.96 × 367 or 3,330 ± 719. Since this interval does not include zero, we can conclude with 95% confidence that the estimated number of male life sciences doctoral recipients is significantly higher than the number of female life sciences doctoral recipients.

However, if there is a high positive (negative) correlation between the two characteristics, the formula will overestimate (underestimate) the true standard error.

In addition to sampling error, data are subject to nonsampling error, which can arise at many points in the survey process. Sources of nonsampling error take many different forms: (1) nonresponse bias, which arises when the characteristics of individuals who do not respond to a survey differ significantly from those who do; (2) measurement error, which arises when we are not able to precisely measure the variables of interest; (3) coverage error, which arises when some members of the target population are not identified and thus do not have a chance to be selected for the sample; and (4) processing error, which can arise at the point of data editing, coding or key entry. These sources of error are much harder to estimate than sampling errors.

Important Notes on the Tables

The following definitions are provided to help facilitate the use of data in the detailed tables.

Field of doctorate is the field of degree as specified by the respondent in the Survey of Earned Doctorates (SED) at the time of degree conferral. These codes were subsequently recoded to the SESTAT codes. (See Appendix A for the doctorate degree fields.)

Occupation data were derived from responses to several questions on the type of work primarily performed by the respondent. The occupational classification of the respondent was based on his/her principal job held during the reference week—or last job held, if not employed in the reference week (questions A20 or A5). Also used in the occupational classification was a respondent-selected job code (questions A21 or A6). (See Appendix B for the list of occupations.)

Sector of employment was based on responses to questions A15 and A17. The category "universities and 4-year colleges" includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), university- affiliated research institutions, and other types of institutions. "Private-for-Profit" includes those self-employed in incorporated business.

Employer location was based primarily on responses to question A11 on the location of the principal employer. Individuals not reporting place of employment were classified by their last mailing address.

Primary work activity was determined from responses to question A30. "Development" includes the development of equipment, products, and systems. "Design" includes the design of equipment, processes, and models.

Federal support was determined from responses to questions A41 and A42.

Faculty rank/tenure status was obtained from the responses to questions A18 and A19.

Race/ethnicity categories of white, black, Asian/ Pacific Islander and American Indian/ Alaskan Native refer to non-Hispanic individuals only. These data are from the SED.

Citizenship status category of non-U.S., temporary resident does not include individuals who, at the time they received their doctorate, expressed plans to leave the United States. These individuals were excluded from the sampling frame.

Salary data were derived from responses to question A34, in which information was requested regarding annual salary before deductions for the principal job held during April 1999, excluding income from bonuses, overtime, and summer teaching/research. Salaries reported are median annual salaries, rounded to the nearest $100 and computed for full-time employed sciences and engineers. For individuals employed by education institutions, no accommodation was made to convert academic-year salaries to calendar-year salaries. Users are advised that due to changes in the salary question since 1993, the 1995, 1997 and 1999 salary data are not strictly comparable with 1993 data.

Labor force participation rate. The labor force is defined as those employed (E) plus those unemployed (U, those not-employed persons actively seeking work). Population (P) is defined as all S&E doctorate holders under age 76, residing in U.S. during the week of April 15, 1999, who earned their doctorate from a U.S. institution. The labor force participation rate (R_{LF}) is the ratio of the labor force to the population (P).
$R_{LF} = (E+U) / P$

Unemployment rate. The unemployment rate (R_u) is the ratio of those who are unemployed but seeking employment (U) to the total labor force (E+U). $R_u = U / (E+U)$

Involuntarily out-of-field rate. The involuntarily out-of-field rate is the percent of employed individuals who reported they were either:

- Working part-time exclusively because suitable full-time work was not available
- Working in an area in their principal job not related to the first doctoral degree at least partially because suitable work in the field was not available.

Summary of Table Changes in 1999 Compared to 1997 Tables

Global changes

1. For all degree field tables, “Computer and information sciences” and “Mathematical sciences” are now separately displayed as broad field groups.
2. Tables were regrouped and renumbered to display the field-of-doctorate-based tables first, followed by the occupation-based tables.
3. Percent distributions were added to most tables in addition to estimated numbers.
4. On all occupation-based tables, “Material/metallurgical engineers” group, which is a larger group, replaced “Industrial engineers.”

Specific table modifications in 1999 [1997 table number]

Table 7 [9,11]	Gender and race/ethnicity tables by doctorate field data are now combined into one table and reported for employed only.
Table 8 [13]	Citizenship status by doctorate field data are reported for employed only.
Table 9 [15]	Age by field of doctorate field data are reported for employed only.
Table 14 [22]	“Primary or secondary” work activity data replace “Primary” work activity.
Table 15 [24]	Puerto Rico is now listed separately from other U.S. Territories.
Table 29 [10,12]	Gender and race/ethnicity by occupation data are reported for employed only.
Table 30 [14]	Citizenship status by occupation data are reported for employed only.
Table 31 [16]	Age by occupation data are reported for employed only.
Table 36 [23]	“Primary or secondary” work activity data replace “Primary” work activity.
Table 37 [25]	Puerto Rico is now listed separately from other U.S. Territories.
Table 40 [30]	“Years since doctorate” data replace “Employer location.”
Table 41 [31,35]	1997 tables 31 and 35 are now combined into one table. “Primary or secondary work activity” and “Years since doctorate” data replace the “Employer location” and “Place of birth” data in 1997 table 31; table title changed to “selected demographic and employment-related characteristics.”

Table 42 [33,34] 1997 tables 33 and 34 are now combined into one table. "Years since doctorate" data replace "Place of birth"; "Primary or secondary work activity" data replace "Primary" work activity; table title changed to "selected demographic and employment-related characteristics."

Table 43 [32,36] 1997 tables 32 and 36 are now combined into one table. "Employment sector" and "Years since doctorate" data replace "Employer location" and "Place of birth"; "Primary or secondary work activity" data replace "Primary" work activity; table title changed to "selected demographic and employment-related characteristics."

Table 48 [57] "Years since doctorate" data replace 'year of doctorate."

Table 53 [58] Puerto Rico is now listed separately from other U.S. Territories.

Table 68 [59] Puerto Rico is now listed separately from other U.S. Territories.

1997 Tables dropped in 1999

1997 table 34 Combined with another table [1997 table 33]

1997 table 35 Combined with another table [1997 table 31]

1997 table 36 Combined with another table [1997 table 32]

1997 tables 48 Median annual salary tables on demographic and employment-related through 56 characteristics

New Tables in 1999

Table 10 Field of doctorate by years since doctorate

Table 12 Number table for 1999 median annual salary table 50 [1997 table 44]

Table 13 Number table for 1999 median annual salary table 51 [1997 table 46]

Table 17 Faculty rank by years since doctorate. "Adjunct" and "Other faculty" data are shown under "All other faculty"

Table 18 Faculty rank by race/ethnicity. "Adjunct" and "Other faculty" data are shown under "All other faculty"

Table 20 Tenure status by years since doctorate

Table 21 Tenure status by race/ethnicity

Table 32 Occupation by years since doctorate

Table 34 Number table for 1999 median annual salary table 65 [1997 table 45]

Table 35 Number table for 1999 median annual salary table 66 [1997 table 47]

Table 46 Median annual salary table for 1999 table 8

Table 47 Median annual salary table for 1999 table 9

Table 54 Median annual salary table for 1999 table 16

Table 55 Median annual salary table for new 1999 table 17

Table 56 Median annual salary table for new 1999 table 18

Table 57 Median annual salary table for 1999 table 19

Table 58 Median annual salary table for new 1999 table 20

Table 59 Median annual salary table for new 1999 table 21

Table 61 Median annual salary table for 1999 table 30

Table 62 Median annual salary table for 1999 table 31

Table 63 Median annual salary table for new 1999 table 32